LSDA - Project 1

Mads Cort Graae, David Mortensen, Oscar Werner

March 2019

0.1 Introduction

The goal of this project is to quantify the performance effects of parallelization. To test these effects, k-means has been implemented in Scala. The running time of the algorithm will be tested on different values of k, to analyze the differences in benefits of paralellization.

0.2 Methods

K-means is a simple unsupervised clustering algorithm. The algorithm starts with randomly initializing k points, which is defined as the cluster centroids.

K-Means is an iterative algorithm and works in many ways similarily to the EM algorithm. It consists of two steps, first the algorithm by calculates the distance from all points to our initialized means, and assigns every point to the nearest centroid. The centroids are then updated by taking the mean of the points assigned to them. These two steps are repeated until convergence.

The algorithm continues

0.3 Amdahl's law

0.4 To do list

- Proper introduction
-
- Plots that compare serial and paralellized version
- Plots to see how running times increase for more clusters - serial and parallelized
-
- Discussion
- Conclusion